Distributional inference

Albers, Casper Johannes

Chapter 6

A problem from the interface between time series and multivariate analysis

> *'Whence, finally, this one thing seems to follow: that if observations of all events were to be continued throughout all eternity, (and hence the ultimate probability would tend toward perfect certainty), everything in the world would be perceived to happen in fixed ratios and according to a constant law of alternation, so that even in the most accidental and fortuitous occurrences we would be bound to recognize, as it were, a certain necessity and, so to speak, a certain fate.*
> *For all I know that is what Plato had in mind when, in the doctrine of the universal circle, he maintained that after the passage of countless centuries everything would return to its original state.'*
>
> JACOB BERNOULLI[1]

The density estimation method of Chapter 3, applied in Chapter 5, refers to situations where an *independent* random sample is considered. In practice, dependencies are often present, because time (or space) is involved, see also the remark in Section 5.5. Almost all theoretical approaches in time series analysis start from stationarity assumptions. If one is studying only *one* time series then it's difficult to think of anything else. In the study of growth curves the statistical perspective of a 'sample' from a 'population' of growth curves is relevant and the stationarity assumption can be dispensed with. In this chapter an issue from the analysis of stationary time series data will be treated by interpreting it as a problem from multivariate statistical analysis. The concept of strong similarity from the theory of distributional inference will be applied. The exact but suboptimal results obtained are contrasted with the asymptotic results some of which obtained earlier by DAVIS ET AL.(1995, 1996, 1998).

6.1 Introduction to the unit root problem

Modern time series analysts tend to remove long-term effects by differencing their time series. At a colloquium in Groningen (April 22th 1999) Richard Davis presented

[1] Jacob Bernoulli, **Ars Conjectandi**, Basileae, 1713; taken from SCHAAFSMA, Laudatio Sir David Cox, *Nieuw Archief voor Wiskunde*, **15**:3, 1997.

a lecture with the title *Recent Developments in the Unit Root Problem for Moving Averages*[2]. Largely summarizing DAVIS ET AL. (1988, 1995, 1996), he studied the following problem. The outcomes $x_1, \ldots, x_n$ of $X_1, \ldots, X_n$ are available and have to be used to test the hypothesis that the time series $\ldots, X_{-1}, X_0, X_1, X_2, \ldots$ satisfies $X_i = E_i - E_{i-1}$, where $\ldots, E_{-1}, E_0, E_1, \ldots$ are independently $\mathcal{N}(0, \sigma^2)$ distributed. As a context, Davis suggested the moving-average model $X_i = E_i + \psi E_{i-1}$, with the assumption $-1 \leq \psi < 1$. This assumption is natural because when $|\psi| < 1$ we have stationarity, a desirable property of a moving average process; for $\psi = -1$ and $\psi = +1$ the distributions of $(X_1, X_2, \ldots, X_n)$ generated 'coincide' (time series analysts call such a MA(1) process with $|\psi| < 1$ an invertible process, and with $|\psi| = 1$ a noninvertible process). Another motivation is from DAVIS AND DUNSMUIR (1996)[3]: 'the maximum likelihood estimator in the $|\theta_0| < 1$ case is asymptotically normal with mean θ_0 and variance $(1 - \theta_0^2)/T$. However, the normal limit provides a particularly inaccurate approximation for values of θ_0 close to ± 1'. More arguments in favour of studying this problem are provided in DAVIS AND DUNSMUIR (1996, p. 3) and DAVIS AND MIKOSCH (1998, p. 100). The last authors claim that the problem is of some interest, because 'for example, a test that a time series has been over-differenced to achieve stationarity is equivalent to testing for the presence of a unit root in the moving average polynomial'. Davis has derived an intriguing asymptotic level-α test, with $\alpha = .045$, by studying the likelihood function $l_x(\theta)$ with respect to the parameter θ which has (ψ, σ^2) as its true value. He established that H_0: $\psi = -1$ should be maintained if and only if $\hat{\psi} = -1$, where $\hat{\psi}$ is the maximum likelihood estimator of ψ (under the restriction $-1 \leq \psi < 1$).

We shall return to Davis' asymptotic approach and other existing methods from literature in Section 6.5. Our interest in the problem has its origin in the relationship with multivariate analysis, especially with that part of multivariate analysis where the attention is concentrated on structured covariance matrices (intra-class correlation, random-effects models, Rao's growth curve models, factor analysis, etcetera).

The problem
Given is the outcome x of $X \sim \mathcal{N}_n(0_n, \Sigma)$ with $\Sigma = (\sigma_{i,j})$ such that

$$\sigma_{i,j} = \begin{cases} (1+\psi^2)\sigma^2 & \text{if } i = j \\ \psi\sigma^2 & \text{if } |i-j| = 1 \\ 0 & \text{if } |i-j| \geq 2 \end{cases}$$

where (ψ, σ^2) is restricted to $[-1, 1) \times (0, \infty)$. Required are

1. a Neyman-Pearson level-α test for $\mathrm{H}_0 : \psi = -1$ against $\mathrm{A} : \psi > -1$;
2. estimates of ψ, σ^2 and, perhaps, some other true values, either in the form of point estimates, or in the form of point estimates with standard errors, confidence intervals or, preferably, distributional inferences.

[2]Transparencies of an earlier lecture with approximately the same content are available from `http://www.stat.colostate.edu/ rdavis/lectures/nber97.pdf`
[3]Davis and Dunsmuir use θ_0 as notation for ψ and T as notation for the sample size.

Note that the correlation matrix (ρ_{ij}) is given by

$$\rho_{i,j} = \begin{cases} \rho = \psi/(1+\psi^2) & \text{if } |i-j| = 1 \\ 0 & \text{if } |i-j| \geq 2 \end{cases}$$

where $\rho \in [-\frac{1}{2}, \frac{1}{2})$ is the first-order auto-correlation. This suggests a parametrization where (ψ, σ^2) is replaced by (ρ, τ^2) with ρ as defined and where $\tau^2 = (1+\psi^2)\sigma^2$ is the variance of X_i $(i = 1, \ldots, n)$. The restrictions now are $(\rho, \tau^2) \in [-\frac{1}{2}, \frac{1}{2}) \times (0, \infty)$ while H_0: $\rho = -\frac{1}{2}$ is the hypothesis of interest.

Yet another parametrization will appear in the context of Section 6.2. Here the parametrization (ν^2, η^2) consists of the two variances

$$\begin{aligned} \nu^2 &= 2\sigma^2(1-\psi+\psi^2) = 2\tau^2(1-\rho) \\ \eta^2 &= 2\sigma^2(1+\psi+\psi^2) = 2\tau^2(1+\rho) \end{aligned}$$

of $X_i - X_{i+1}$ and $X_i + X_{i+1}$, respectively. Here the restrictions are

$$\frac{1}{3} \leq \frac{\eta^2}{\nu^2} = \frac{1+\rho}{1-\rho} < 3$$

and H_0: $\eta^2/\nu^2 = \frac{1}{3}$ is the null hypothesis to be tested.

Remark 1

The restrictions $\psi \in [-1, 1)$, $\rho \in [-\frac{1}{2}, \frac{1}{2})$ and $\eta^2/\nu^2 \in [\frac{1}{3}, 3)$ have been imposed to make the parameters identifiable in the sense that if X is generated as prescribed and its distribution is given, then the parameter is uniquely determined. Note that the restriction $-1 \leq \psi < 1$ has its origin in the idea that the contribution of E_i should be 'more important' than that of E_{i-1} in $X_i = E_i + \psi E_{i-1}$ (in the sense that $\mathrm{Var}\, E_i \geq \mathrm{Var}\, (\psi E_{i-1})$). Starting from $-1 \leq \psi < 1$, the correlation $\rho = \psi(1+\psi^2)^{-1}$ is in $[-\frac{1}{2}, \frac{1}{2})$. The inverse transformation is

$$\psi = \frac{1 - \sqrt{1-4\rho^2}}{2\rho}$$

(The other root, $(1 + \sqrt{1-4\rho^2})/2\rho$, provides values beyond the interval $[-1, 1)$ if $\rho \in [-\frac{1}{2}, \frac{1}{2})$.)

Remark 2

The situation is also of general methodological interest because the parametrizations with

$$\begin{array}{lll} (\psi, \sigma^2) & \text{as the true value of} & \theta^{(1)} = \left(\theta_1^{(1)}, \theta_2^{(1)}\right) \in \Theta^{(1)} \\ (\rho, \tau^2) & \text{as the true value of} & \theta^{(2)} = \left(\theta_1^{(2)}, \theta_2^{(2)}\right) \in \Theta^{(2)} \\ (\nu^2, \eta^2) & \text{as the true value of} & \theta^{(3)} = \left(\theta_1^{(3)}, \theta_2^{(3)}\right) \in \Theta^{(3)} \end{array}$$

are all well-motivated. Some methods of inference depend on the parametrization. The situation will then require a specific analysis. The domains $\Theta^{(1)}$, $\Theta^{(2)}$, and $\Theta^{(3)}$

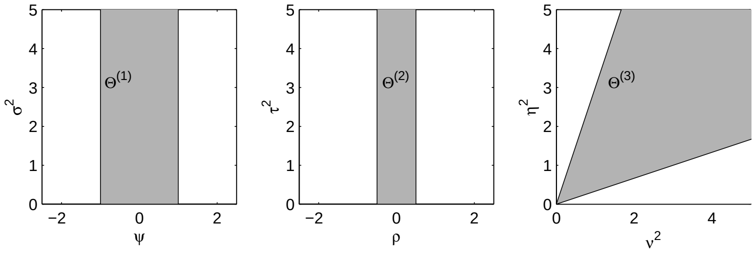

Figure 6.1: Visualization of the restricted area for each parametrization

are visualized in Figure 6.1. Other parametrizations, e.g. $\theta^{(4)} = (\eta^2/\nu^2, \tau^2)$, are of course also possible.

Preview

The case $n = 2$ to be discussed in Section 6.2 is of methodological but of no practical interest because nobody should consider the testing of H_0 or the estimation of parameters if only 2 outcomes are available. More than a few observations are necessary before it is reasonable to make any statistical inference about the true value t of any θ. Section 6.3 is about $n \geq 3$. The methodological aspects will not be treated as satisfactorily as in Section 6.2 but the exact methods to be developed are of practical interest, especially if n is neither too small, nor too large (because the normality assumption becomes questionable if n is very large while, moreover, asymptotic methods can be incorporated).

6.2 The case $n = 2$

Testing

To test H_0 in case $n = 2$, the parametrization with (ν^2, η^2) as the true value of $\theta^{(3)} \in \Theta^{(3)}$ is most convenient because the $1:1$ transformation

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 - X_2 \\ X_1 + X_2 \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \nu^2 & 0 \\ 0 & \eta^2 \end{bmatrix} \right)$$

makes it obvious that H_0: $\eta^2/\nu^2 = \frac{1}{3}$ should be tested against A: $\eta^2/\nu^2 > \frac{1}{3}$ by rejecting H_0 iff $3Y_2^2/Y_1^2 \geq F_{1,1;\alpha}$. Note that (Y_1^2, Y_2^2) is a complete sufficient statistic in the general situation and that $3Y_1^2 + Y_2^2$ is a complete sufficient statistic for the null hypothesis. The theory of multi-parameter exponential families (LEHMANN, 1986) provides that the F-test discussed is UMP invariant level α (p. 290), UMP similar level α (p. 199), and UMP unbiased level α (p. 122).

Estimating

The first part of the problem being solved, we can discuss the second part, the estimation of various parameters, preferably by means of confidence intervals or distribution functions of distributional inferences.

Point estimates

Continuing with the third parametrization, we have that Y_1^2 and Y_2^2 are best unbiased estimates of ν^2 and η^2. As $F_{1,1} \sim \text{Cauchy}^2$ does not have a finite expectation, η^2/ν^2 does not allow an unbiased estimate. The maximum likelihood method is complicated by the fact that Y_1^2 and Y_2^2 are *not* necessarily ML estimates of ν^2 and η^2 because the restriction $\frac{1}{3} \leq \eta^2/\nu^2 < \frac{1}{3}$ has to be complied with. The ML estimates $\hat{\eta}^2, \hat{\nu}^2$ are such that

$$\begin{array}{lll} y_2^2/y_1^2 \leq \frac{1}{3} & \text{implies } \hat{\eta}^2/\hat{\nu}^2 = \frac{1}{3} & (\text{in fact } \hat{\nu}^2 = \frac{1}{6}(y_1^2 + 3y_2^2),\ \hat{\eta}^2 = \frac{1}{3}\hat{\nu}^2) \\ \frac{1}{3} \leq y_2^2/y_1^2 \leq 3 & \text{implies } \hat{\eta}^2/\hat{\nu}^2 = y_2^2/y_1^2 & (\text{in fact } \hat{\nu}^2 = y_1^2,\ \hat{\eta}^2 = y_2^2) \\ y_2^2/y_1^2 \geq 3 & \text{implies } \hat{\eta}^2/\hat{\nu}^2 = 3 & (\text{in fact } \hat{\nu}^2 = \frac{1}{6}(3y_1^2 + y_2^2),\ \hat{\eta}^2 = 3\hat{\nu}^2) \end{array}$$

(where, in principle, the latter value is not allowed but this complication is of no practical interest).

Since $(1+\rho)/(1-\rho) = \eta^2/\nu^2$ we have that

$$\hat{\rho} = \begin{cases} -\frac{1}{2} & \text{if } r \leq -\frac{1}{2} \\ r & \text{if } -\frac{1}{2} \leq r \leq \frac{1}{2} \\ \frac{1}{2} & \text{if } r \geq \frac{1}{2} \end{cases}$$

where $r = 2x_1x_2/(x_1^2 + x_2^2)$ can vary between -1 and $+1$. From the relation $\tau^2 = \frac{1}{4}(\nu^2 + \eta^2)$ we directly obtain the ML estimate

$$\hat{\tau}^2 = \begin{cases} (x_1 + x_2)^2 & \text{if } y_2^2/y_1^2 \leq \frac{1}{3} \\ \frac{1}{2}(x_1^2 + x_2^2) & \text{if } \frac{1}{3} \leq y_2^2/y_1^2 \leq 3 \\ (x_1 - x_2)^2 & \text{if } y_2^2/y_1^2 \geq 3 \end{cases} .$$

Similarly, the relationship $\rho = \psi/(1+\psi^2)$ and $\tau^2 = (1+\psi^2)\sigma^2$ can be used in constructing ML estimates of (ψ, σ^2). This method is appealing to generate point estimates, but to express statistical uncertainties, other methods are needed.

Bayesian methods (likelihood inference included)

Given some parametrization we can discuss the likelihood function, incorporate a weight function w and derive the posterior density

$$q(\theta_1, \theta_2) = \frac{L_x(\theta_1, \theta_2) w(\theta_1, \theta_2)}{\int\int_{\Theta} L_x(\theta_1, \theta_2) w(\theta_1, \theta_2)\, \mathrm{d}\theta_1\, \mathrm{d}\theta_2}.$$

Choosing the Jeffreys prior $w(\theta) = |I(\theta)|^{1/2}$ is of some interest because it implies that the posterior densities are probabilistically coherent (JEFFREYS, 1932, and 1939, Ch. 3). Although Jeffreys's prior is widely accepted amongst Bayesians for single parameter models, in multi-parameter models its is more controversial (see, e.g., DATTA ET AL., 2000). At a deeper level even this choice can be criticized as Jeffreys himself seems to have done. ‘In almost all cases we can approach this [Jeffreys] prior as the limit of a sequence of proper (normalizable) priors, with mathematically well-behaved results. If even that does not yield a proper posterior distribution, then the robot is warning us that the data are too uninformative about either very large s or very small s to justify any definite conclusions, and we need to get more evidence before any useful inferences are possible' (JAYNES, 1996, p. 629).

For the parametrization $\theta^{(3)} = (\nu^2, \eta^2)$ we obtain the likelihood function

$$L_x^{(3)}(\theta_1, \theta_2) = \frac{1}{2\pi\sqrt{\theta_1\theta_2}} \exp\left[-\frac{1}{2}\left(\frac{(x_1 - x_2)^2}{\theta_1} + \frac{(x_1 + x_2)^2}{\theta_2}\right)\right].$$

and the information matrix

$$I_x^{(3)}(\theta_1, \theta_2) = \begin{bmatrix} \frac{1}{2\theta_1^2} & 0 \\ 0 & \frac{1}{2\theta_2^2} \end{bmatrix}.$$

Hence $|I_x^{(3)}(\theta_1, \theta_2)|^{1/2} = 1/(2\theta_1\theta_2)$ and

$$q^{(3)}(\theta_1, \theta_2) = C\theta_1^{-3/2}\theta_2^{-3/2} \exp\left(-\frac{1}{2}\left(\frac{(x_1 - x_2)^2}{\theta_1} + \frac{(x_1 + x_2)^2}{\theta_2}\right)\right)$$

if $\theta_2/\theta_1 \in [\frac{1}{3}, 3)$, with the proportionality constant C obtained via integrating over $\Theta^{(3)}$, yielding

$$\begin{aligned} C^{-1} &= \int_0^\infty \int_{\theta_1/3}^{3\theta_2} L_x^{(3)}(\theta_1, \theta_2) w^{(3)}(\theta_1, \theta_2)\, \mathrm{d}\theta_2 \mathrm{d}\theta_1 \\ &= \frac{4}{\sqrt{|x_1^2 - x_2^2|}} \left(\arctan\sqrt{\frac{3(x_1 + x_2)^2}{(x_1 - x_2)^2}} - \arctan\sqrt{\frac{(x_1 + x_2)^2}{3(x_1 - x_2)^2}}\right). \end{aligned}$$

The information matrices, and hence Jeffreys likelihood inferences, for other parametrizations can either be obtained directly or by using Jacobian matrices. This yields

$$|I_x^{(1)}(\theta_1, \theta_2)|^{1/2} = \frac{1 - \psi^2}{\sigma^2(1 + \psi^2 + \psi^4)} \quad \text{and} \quad |I_x^{(2)}(\theta_1, \theta_2)|^{1/2} = \frac{1}{(1 - \rho^2)\tau^2}.$$

Although results can be derived, we are not satisfied. Furthermore, JAYNES (1996, p. 630) noted that 'while the Jeffreys prior is theoretically the correct one, it is in practice a small refinement that makes a difference only in the very small sample case'. We do not agree with the qualification as 'the' correct one because integration with respect to Lebesgue measure is only one possibility. Moreover no prior probability mass is assigned to the hypothesis H_0: $\psi = -1$ of interest.

Fisher-Neyman compromise

Our main interest is in constructing distributional inferences for η^2/ν^2, ρ and ψ using a modernization of Fisher's fiducial argument. The idea is very simple: one-sided P-values are used to express the degree of belief in one-sided hypotheses (SALOMÉ, 1998; see also FISHER, 1973, p. 63, where he, in some sense, discredits his unpolished version this useful idea). Under certain restrictive conditions, 'optimality' of thus defined procedures can be established (they are 'best strongly similar', see KROESE ET AL., 1999). The analogy with 'best unbiased estimates' is obvious. The modernized fiducial argument is seldomly applicable and if it is applicable then the methods of inference generated may still fail to be satisfactory. Much of the work done in asymptotic theory can be regarded as in line with this modernized fiducial argument. One is establishing that methods of inference are 'asymptotically strongly similar' in the sense that 'confidence intervals based on it are asymptotically correct', etcetera. The 'exactness' of the fiducial method is very appealing. Consider the problem of making a distributional inference about η^2/ν^2. We need a procedure $Q : [0,\infty)^2 \to [\frac{1}{3}, 3)^*$ (the space of probability distributions on $[\frac{1}{3}, 3)$) such that $Q(x)$ provides our 'opinion' about the true value η^2/ν^2 of $\theta_2^{(3)}/\theta_1^{(3)}$. Let $\mathcal{G} = \{G_x |\ x \in (0,\infty)^2\}$ denote the family of distribution functions. With $G_x(z)$ we express the degree of belief in H_z: $\eta^2/\nu^2 \le z$. We have, of course, $G_x(z) = 0$ if $z \le \frac{1}{3}$ and $G_x(z) = 1$ if $z \ge \frac{1}{3}$, and the Fisher-Neyman compromise prescribes to define $G_x(z) = \alpha_z(x)$ where $\alpha_z(x)$ is the P-value if H_z: $\eta^2/\nu^2 \le z$ is tested against A_z: $\eta^2/\nu^2 > z$. This, of course, is done with the F-test using Y_2^2/Y_1^2 as test statistic having the $zF_{1,1}$ distribution if $\eta^2 = z\nu^2$. We simply have

$$\alpha_z(x) = \mathrm{P}\left(F_{1,1} \ge \frac{y_2^2}{zy_1^2}\right) = \mathrm{P}\left(F_{1,1} \le z\frac{y_1^2}{y_2^2}\right),$$

because $\mathcal{L}\,(F_{1,1}) = \mathcal{L}\,(F_{1,1}^{-1})$ and have

$$G_x(z) = \begin{cases} 0 & \text{if } z < \frac{1}{3} \\ \mathrm{P}\left(F_{1,1} \le z\frac{y_1^2}{y_2^2}\right) = \frac{2}{\pi}\arctan\sqrt{z\frac{y_1^2}{y_2^2}} & \text{if } \frac{1}{3} < z < 3 \\ 1 & \text{if } z \ge 3 \end{cases}$$

as our distributional inference[4]. This family is *strongly similar* in the sense that $G_X^*(z) \sim \mathcal{U}(0,1)$ holds if $\eta^2/\nu^2 = z$ (except in the case $z = 3$ where a degeneracy appears). Note that this distributional result implies that

$$\mathrm{P}\left(G_X^{-1}\left(\frac{1}{2}\alpha\right) \le \frac{\eta^2}{\nu^2} \le G_X^{-1}\left(1 - \frac{1}{2}\alpha\right)\right) > 1 - \alpha$$

(strict inequality is caused by the degeneracies at $\frac{1}{3}$ and 3). It can be established along the lines of KROESE ET AL. (1999) that any other family $\{G_x; x = (x_1, x_2) \in (0,\infty)^2\}$ satisfying the strong similarity condition is 'uniformly worse' than $\{G_x^*; x = (x_1, x_2) \in (0,\infty)^2\}$ from a variety of perspectives. (The degeneracies at $\frac{1}{3}$ and 3 have the effect that the conditions used in KROESE ET AL. (1999) are not satisfied. The statement

[4]The result for $\frac{1}{3} < z < 3$ is easily seen after observing that

$$f_{1,1}(u) = \frac{1}{\pi(1+u)\sqrt{u}} \quad \to \quad F_{1,1}(u) = \int_0^u \frac{\mathrm{d}x}{\pi(1+v)\sqrt{v}} = \frac{2}{\pi}\arctan\sqrt{u}.$$

is not completely correct. That is why we used ‘ ’). Above ‘fiducial inference’ can immediately be transformed to obtain optimal distributional inferences for

$$\rho = \frac{\eta^2 - \nu^2}{\eta^2 + \nu^2} \quad \text{and} \quad \psi = \frac{1 - \sqrt{1 - 4\rho^2}}{2\rho}.$$

Conclusion
The case $n = 2$ is, as said before, only of methodological interest. The Neyman-Pearson test for H_0 and distributional inferences for η^2/ν^2, ρ, and ψ are fairly attractive though, of course, nobody should consider such problems seriously if n is only 2. The restrictions to the parameter space cause trouble if Bayesian inferences (with respect to continuous weight functions w) are considered. Maximum likelihood estimators are quite attractive, though solely for the purpose of point estimation.

6.3 Exact inferences in the case $n > 2$

For the parametrization (ν^2, η^2), just obtained the distributional result

$$\frac{Y_2^2}{Y_1^2} \sim \frac{1+\rho}{1-\rho} F_{1,1}$$

as the basis of *exact* inferences. A situation of more interest is that where observable random vectors $(Y_{1,1}, Y_{1,2}), \ldots, (Y_{m,1}, Y_{m,2})$ are independently distributed with the same distribution as (Y_1, Y_2). As a basis of the inference one will use the outcome v of

$$V = \frac{\sum_{j=1}^m Y_{2,j}^2}{\sum_{j=1}^m Y_{1,j}^2} \sim \frac{1+\rho}{1-\rho} F_{m,m}.$$

This solution can be enforced to our actual situation by defining $m = \lfloor \frac{n+1}{3} \rfloor$, and by taking

$$\begin{pmatrix} Y_{1,j} \\ Y_{2,j} \end{pmatrix} = \begin{pmatrix} X_{3j-1} - X_{3j-2} \\ X_{3j-1} + X_{3j-2} \end{pmatrix}$$

($j = 1, \ldots, m$, tacitly skipping the observations $X_3, X_6, \ldots$). One possibility to test $\mathrm{H}_0 : \psi = -1$ (or, equivalently, $\rho = -\frac{1}{2}$) against $\mathrm{H}_1 : \psi > -1$ at level α is to reject H_0 if and only if $3V \geq F_{m,m;\alpha}$. This test is UMP unbiased size α, UMP similar size α, and UMP invariant size α (see, again, LEHMANN, 1986); but, of course, only after the deletion of $X_3, X_6, \ldots$.

The results thus obtained are exact, yet not satisfactory because one third of the data has been ignored. We can develop alternative approaches with

$$\begin{aligned} V^* &= \frac{\sum_{j=1}^m (X_{3j} + X_{3j-1})^2}{\sum_{j=1}^m (X_{3j} - X_{3j-1})^2} \sim \frac{1+\rho}{1-\rho} F_{m,m} \\ V^{**} &= \frac{\sum_{j=1}^m (X_{3j+1} + X_{3j})^2}{\sum_{j=1}^m (X_{3j+1} - X_{3j})^2} \sim \frac{1+\rho}{1-\rho} F_{m,m} \end{aligned}$$

based on the deletion of $X_1, X_4, \ldots$ and $X_2, X_5, \ldots$ respectively (assuming, for the sake of simplicity, that $n = 3m + 1$). The construction of tests, confidence intervals, and distributional inferences using V^* or V^{**} is identical to that using V, since the three have the same distribution. However, they are not independent.

Bonferroni-inequality based combinations can be obtained, but these are too 'conservative'. We would like to do better. One intuitive approach consists of combining V, V^*, V^{**} by adding the numerators and denominators separately. This provides the statistic

$$W = \frac{\sum_{i=1}^{n-1}(X_{i+1} + X_i)^2}{\sum_{i=1}^{n-1}(X_{i+1} - X_i)^2} = \frac{1+R}{1-R}$$

where

$$R = \frac{2\sum_{i=1}^{n-1} X_i X_{i+1}}{\sum_{i=1}^{n-1}(X_i^2 + X_{i+1}^2)}$$

is a natural estimator of the autocorrelation ρ. Asymptotic theory can be developed. These intuitive manipulations, however, are not completely satisfactory either; why should the combination of V, V^*, V^{**} be done by adding numerators and denominators? In this respect, work on combining expert opinions is of some interest (cf. GENEST AND ZIDEK, 1986, GENEST ET AL., 1990). In the remainder of this chapter various attempts are described to obtain more satisfactory 'asymptotic' results. The discussion will be somewhat confusing. We are not satisfied by it. In our opinion the exact results just described (F test for H_0, strongly similar distributional inferences for $(1+\rho)/(1-\rho)$, ρ, ψ) are attractive in spite of the fact that one third of the observations is discarded from the analysis. (The loss of information is probably less because of the correlation.)

6.4 Focusing on exact tests for the null hypothesis

In his breakthrough paper, KARL PEARSON (1900) focused on the null hypothesis that certain theoretical probabilities were correct. Later generations of statisticians emphasized that the 'set of all alternatives' should be taken into account. From here we restrict the attention to the testing of H_0:$\psi = -1$ i.e. $X_i = E_i - E_{i-1}$ $(i = 1, \ldots, n)$ where the E_i are i.i.d. $\mathcal{N}(0, \sigma^2)$. This has as a consequence that the suggestions in the last part of Section 6.3 are not necessarily useful. A different suggestion is as follows. As taking partial sums is the inverse of differencing, we consider

$$\begin{aligned} S &= \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_n \end{pmatrix} = \begin{pmatrix} X_1 \\ X_1 + X_2 \\ \vdots \\ X_1 + \ldots + X_n \end{pmatrix} \overset{\mathrm{H}_0}{=} \begin{pmatrix} E_1 - E_0 \\ E_2 - E_0 \\ \vdots \\ E_n - E_0 \end{pmatrix} \\ &\sim \mathcal{N}_n\left(0_n, \sigma^2(I_n + \iota_n \iota_n')\right). \end{aligned}$$

(ι_n is the vector of size n with all entries one). The largest eigenvalue of $I_n + \iota_n\iota_n'$ is $n+1$ with corresponding eigenvector $n^{-1/2}\iota_n$. All other eigenvalues are equal to 1, the corresponding eigenvectors can be chosen arbitrarily as long as they are perpendicular and have elements that sum up to zero. To fix a choice, we use the matrix

$$U_n = \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\sqrt{n^2-n}} & \frac{1}{\sqrt{n^2-n}} & \frac{1}{\sqrt{n^2-n}} & \cdots & \frac{-(n-1)}{\sqrt{n^2-n}} \end{bmatrix}$$

(HELMERT, 1876). We now have that $I_n + \iota_n\iota_n' = U_n'\Lambda_n U_n$ where $\Lambda_n = \mathrm{diag}(n+1, 1, \ldots, 1)$. Hence $\mathcal{L}\ (U_nS) = \mathcal{N}_n(0_n, \sigma^2\Lambda_n)$ and $T = \Lambda_n^{-1/2}U_nS \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$. Note that there is a $1:1$ correspondence between T and X. Now it is obvious that a large number of F-tests can be constructed. For each $f, g \in \mathbb{N}$, with $f + g \leq n$

$$\frac{(T_1^2 + \ldots + T_f^2)/f}{(T_{f+1}^2 + \ldots + T_{f+g}^2)/g} \sim F_{f,g}$$

provides a test. At first sight, $f = 1, g = n-1$ might be of particular interest. In that case, elementary computations provide that $T_1^2 = \frac{n}{n+1}\bar{S}^2$ and $T_2^2 + \ldots + T_n^2 = \sum_{i=1}^n (S_i - \bar{S})^2$, where $\bar{S}$ is the average of the partial sums. But, eventually, a more even partitioning of $n = f + g$ into f and g may be more appropriate. The message is clear: the alternative hypothesis should play its part.

Testing against specific alternatives

It follows from the above that the null hypothesis allows the complete sufficient statistic $\|T\|_2^2 \sim \sigma^2\chi_n^2$. Hence, given a specific alternative (ψ, σ^2) with $\psi > -1$ (or (ρ, τ^2) with $\rho > -\frac{1}{2}$), we can try to obtain the MP similar size-α test by using the theory of tests with Neyman structure. For $n = 2$ the test suggested before is UMP similar, unbiased size-α. For $n = 3$ the situation is already very complicated because the model does not allow a 2-dimensional exponential family representation: a 3-dimensional family is needed. In case $n = 3$ (later on, results will be generalized to the case of general n) we have that

$$T = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} \psi & 1 & 0 & 0 \\ \psi & 1+\psi & 1 & 0 \\ \psi & 1+\psi & 1+\psi & 1 \end{bmatrix} \begin{pmatrix} E_0 \\ E_1 \\ E_2 \\ E_3 \end{pmatrix}$$

is such that $T \sim \mathcal{N}_3(0_3, \sigma^2\Sigma_\psi)$ where

$$\Sigma_\psi = \Lambda_n^{-1/2}U_n \begin{bmatrix} \psi & 1 & 0 & 0 \\ \psi & 1+\psi & 1 & 0 \\ \psi & 1+\psi & 1+\psi & 1 \end{bmatrix} \begin{bmatrix} \psi & \psi & \psi \\ 1 & 1+\psi & 1+\psi \\ 0 & 1 & 1+\psi \\ 0 & 0 & 1 \end{bmatrix} U_n^{-1}\Lambda_n^{-1/2}$$

can, in principle, be computed for any fixed $\psi \in [-1, 1)$. (Note that Σ_{-1} is equal to the identity matrix, as we saw earlier.) Let V be an orthogonal matrix such that $\Sigma_\psi = V\widetilde{\Lambda_n}V'$ where $\widetilde{\Lambda_n} = \text{diag}(\lambda_1, \lambda_2, \lambda_3)$. Note that V is a known matrix (if ψ is given). In terms of $\tilde{T} = VT$ the problem of determining the test with Neyman-structure, optimal in (ψ, σ^2) is that where the conditional distribution of $\tilde{T}$, given the value of $\|\tilde{T}\|_2^2$ observed, has to be considered under the alternative hypothesis that the $\tilde{T}_i \sim \mathcal{N}(0, \sigma^2\lambda_i)$ are independently distributed. As

$$\mathcal{L}\,(\tilde{T}) = (\lambda_1\lambda_2\lambda_3)^{-1/2}\left(2\pi\sigma^2\right)^{-3/2}\exp\left[-\frac{1}{2\sigma^2}\left(\frac{t_1^2}{\lambda_1}+\frac{t_2^2}{\lambda_2}+\frac{t_2^3}{\lambda_3}\right)\right]$$

we shall have to reject H_0 for small values of $\lambda_1^{-1}t_1^2+\lambda_2^{-1}t_2^2+\lambda_3^{-1}t_3^2$ given the outcome $\tilde{T}_1^2+\tilde{T}_2^2+\tilde{T}_3^2$.

The critical level has to be determined by studying the conditional distribution of $\lambda_1^{-1}\tilde{T}_1^2 + \lambda_2^{-1}\tilde{T}_2^2 + \lambda_3^{-1}\tilde{T}_3^2$ given $\|\tilde{T}\|_2^2$ (under H_0). This conditional distribution does not depend on σ^2 because $\|\tilde{T}\|_2^2 \sim \sigma^2\chi_3^2$ is a (complete) sufficient statistic. One can equally well study the distribution of

$$\left(\frac{\tilde{T}_1^2}{\lambda_1}+\frac{\tilde{T}_2^2}{\lambda_2}+\frac{\tilde{T}_3^2}{\lambda_3}\right)/\|\tilde{T}\|_2^2$$

because $\left(\frac{\tilde{T}_1^2}{\|\tilde{T}\|_2^2}, \frac{\tilde{T}_2^2}{\|\tilde{T}\|_2^{2}}^{2}, \frac{\tilde{T}_3^2}{\|\tilde{T}\|_2^2}\right)$ and $\|\tilde{T}\|_2^2$ are independent (under H_0), this follows from Basu's theorem. Moreover,

$$\mathcal{L}\left(\frac{\tilde{T}_1^2}{\|\tilde{T}\|_2^2}, \frac{\tilde{T}_2^2}{\|\tilde{T}\|_2^2}, \frac{\tilde{T}_3^2}{\|\tilde{T}\|_2^2}\right)$$

is the Dirichlet$(\frac{1}{2}, \frac{1}{2}; \frac{1}{2})$ distribution (see, e.g., WILKS, Section 7.7). For general n, this result extends to the Dirichlet$(\frac{1}{2}, \ldots, \frac{1}{2}; \frac{1}{2})$ distribution and corresponding test statistic. Elaboration of this extension is not made because the argumentation is an analogue of that for the case $n = 3$ given above, while the choice of ψ and that of the λ_i remains problematic.

6.5 Survey of the literature for the case $n = \infty$

In this section we will review existing asymptotic theory on the unit root problem.

DAVIS ET AL. (1995) and DAVIS AND DUNSMUIR (1996), amongst others, discuss the MA(1) process $y_t = \epsilon_t - \theta_0\epsilon_{t-1}$ where the $\epsilon \sim$ i.i.d.$(0, \sigma^2)$ are not necessarily Gaussian. Both articles consider the reparametrization $\theta_T = 1 - \beta/T$ where $\beta \geq 0$ and T is the sample size. The true parameter is denoted by $\theta_0 = 1 - \gamma/T$. This idea of a value of $\psi = -\theta_0 = -1 + \gamma/n$ very close to -1 may provide us with a *locally* best similar size-α test if we incorporate the idea in Section 6.4. This is an interesting subject for further research.

DAVIS ET AL. (1995) discuss four different tests for H: $|\theta_0| = 1$ versus H: $|\theta_0| < 1$. The first one is based on $\hat{\theta}_{\text{MLE}}$, 'the value of θ which maximizes the likelihood over the interval $\theta \in [-1, 1]$', the second one on $\hat{\theta}_{\text{LM}}$, 'defined as the local maximizer of the likelihood closest to $\theta = \pm 1$'. The asymptotic distributions of $\hat{\theta}_{\text{MLE}}$ and $\hat{\theta}_{\text{LM}}$ are *not* the same. The authors prefer the Local Maximizer estimator for slightly better behavior and easier computation. The third and fourth test are the generalized likelihood ratio and Tanaka's score type test (TANAKA, 1990), respectively. DAVIS ET AL. (1995) state that 'since 'the distribution of the MLE of θ under H_0: $\theta = 1$ is unknown even asymptotically,' TANAKA (1990) was 'reluctant to use such tests as likelihood ratio test of Wald tests'.

Simulations suggest the use of the third or fourth test. Other simulations provide critical values $b_{\text{MLE}}(\alpha)$ and $b_{\text{LM}}(\alpha)$ for the Maximum Likelihood and Local Maximum estimators for β, and hence, yield tests for H_0, in case the ϵ are normally distributed. The limiting behavior of $\hat{\theta}_{\text{LM}}$ under a sequence of local values of θ_0 which converge to 1 at rate $1/T$ is considered. 'In particular, [...] $T(1 - \hat{\theta}_{\text{LM}}) \to^d \beta_0$ where β_0 is the minimizer of some stochastic process. The limit distribution has both a discrete component at the value 0, called the pile-up effect, and a continuous component' (DAVIS AND MIKOSCH, 1998, p. 100). This pile-up effect corresponds to the probability that, under the hypothesis, the LM estimator is equal to 1. For example for, again, $T = 50$ and $\alpha = 5\%$, we have P_{H_0} $(\hat{\theta}_{\text{LM}} = 1) = .649$ (exact result, DAVIS AND DUNSMUIR, 1996, p. 24). Furthermore, on basis of simulations, tests for H_0 for various levels α can be constructed. For example, when $T = 50$ and $\alpha = 5\%$, H_0 is rejected iff $\hat{\theta}_{\text{LM}} < .874$. DAVIS ET AL. (lecture notes mentioned in Footnote 1, p. 17) provide P_{H_0} $(\hat{\theta}_{\text{MLE}} = 1) = .955$. Hence, an exact level $\alpha = 0.045$ test for H_0, based on the maximum likelihood is to reject H_0 iff the MLE differs from 1.

DAVIS AND MIKOSCH discuss the MA(1) process $y_t = \epsilon_t - \theta_0 \epsilon_{t-1}$ where $|\theta_0| \leq 1$ and $\{\epsilon_t\}$ is an i.i.d. sequence of symmetric α-stable variables (see, e.g., BILLINGSLEY (1995, p. 378)) which implies that the process has infinite variance. Here the pile-up effect seems to be slightly smaller than in the finite variance case.

6.6 Discussion

The applied statistician is confronted with many challenging problems. In practice independence assumptions are very often not satisfied because time and/or space dependencies are involved. The theory of *exact* distributional inferences (or of exact confidence intervals) is often illustrated in classroom examples with i.i.d. observations. That is one of the reasons why we were challenged by Davis' lecture. One other reason is that his problem belonged to multivariate analysis with structured covariance matrices allowing an 'exact' treatment. The 'exact' results obtained (for small n) can be compared with the asymptotic results of Davis and others.